Proceedings of the CLTW 4 @ LREC2022 , pages 47–51
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Automatic Speech Recognition for Irish: the ABAIR-ÉIST System

Liam Lonergan[1], Mengjie Qian[2], Harald Berthelsen[1], Andrew Murphy[1], Christoph Wendler[1], Neasa Ní Chiaráin[1], Christer Gobl[1], Ailbhe Ní Chasaide[1]
[1]Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin,
[2]Department of Engineering, Cambridge University
{llonerga, berthelh, murpha61,wendlec,nichiarn,cegobl,nichsid}@tcd.ie
mq227@cam.ac.uk

Abstract

This paper describes ÉIST, automatic speech recogniser for Irish, developed as part of the ongoing ABAIR initiative, combining (1) acoustic models, (2) pronunciation lexicons and (3) language models into a hybrid system. A priority for now is a system that can deal with the multiple diverse native-speaker dialects. Consequently, (1) was built using predominately native-speaker speech, which included earlier recordings used for synthesis development as well as more diverse recordings obtained using the MíleGlór platform. The pronunciation variation across the dialects is a particular challenge in the development of (2) and is explored by testing both Trans-dialect and Multi-dialect letter-to-sound rules. Two approaches to language modelling (3) are used in the hybrid system, a simple n-gram model and recurrent neural network lattice rescoring, the latter garnering impressive performance improvements. The system is evaluated using a test set that is comprised of both native and non-native speakers, which allows for some inferences to be made on the performance of the system on both cohorts.

Keywords: Irish, speech recognition, minority language

1. Introduction

This paper describes the ongoing work to develop automatic speech recognition (ASR) systems for Irish, as part of the ABAIR initiative on Irish speech technology. The current system, ÉIST is described, and the results of recent tests are presented, along with the pointers to issues that are pertinent to all Celtic (and other endangered) languages

2. Background

The development of automatic speech recognition (ASR) for Irish, is a current goal in the ABAIR ("to speak") research programme at the Phonetics and Speech Laboratory, Trinity College Dublin. ABAIR is concerned with developing speech technologies for Irish as well as applications that make the technology useful for the language community. Fundamental to this work is the provision of the linguistic research, that not only delivers resources that underpin the technology, but is also essential to the wider language research community and central to the building of 'intelligent' applications – i.e, applications that incorporate a knowledge of the language structure. Lingustic resource building has been a central feature of our work from the outset – from the original collaboration between Irish and Welsh researchers to develop speech resources for the two languages (the EU-Interreg project WISPR[1]).

In developing speech technology for Irish, and similarly for other Celtic (and endangered languages) one needs to consider that there is no one spoken standard – but rather three dialects which diverge considerably in lexicon, morphology, and especially in pronunciation. Whereas, in the 'major' widely spoken languages, technologies were developed for a standard variety (catering for other varieties came much later) this was/is not an option for Irish. Thus, developing text-to-speech synthesis was approached from the outset as a multi-dialect project, requiring the development of linguistic resources that could provide for a multi-dialect facility. Text-to-speech synthesis systems have been developed for the three main dialects of Irish: Ulster (Ul); Connacht (Co) and Munster (Mu). The synthetic voices, which are available on the ABAIR website[2] include male and female voices and the user has a choice of speech engines (currently deep neural network (DNN) and hidden Markov model speech synthesis (HTS) voices). Current work is focussed on extending the range of dialects covered, as well as exploring the rapidly evolving synthesis modalities.

In building core technology, such as speech synthesis or recognition, it makes sense to understand (i) what applications are most needed by the language community and (ii) who precisely the users might be. To date, ABAIR has been exploiting the synthetic voices in applications for (i) the *general public*, e.g., a web-reader that reads out any electronic text in your choice of dialect; (ii) for Irish language *teaching and learning*, e.g. learning platforms geared to different learner cohorts, different language skills and different language levels (Ní Chiaráin et al., 2022), and (iii) for *disability and access*, to enable the inclusion of this very neglected 'minority within the minority' (Barnes et al., 2022).

In building the ASR system, the diversity of the potential users/applications presents many challenges. As in the development of speech synthesis systems, it is a basic requirement that the system can deal equally well with the diversity of native dialects. Furthermore, one envisages many applications in the educational sphere, where recognition of learners' productions is desirable. This latter group is in itself a very diverse cohort – with different levels ranging from highly proficient speakers with near native-speaker pronunciation to beginners, to fluent speakers who have, nonetheless, a sound system more akin to that of English. Furthermore, for educational and disability applications, one will need recognition of children's voices.

[1] https://keep.eu/projects/2540/WISPR-EN/

[2] https://abair.ie

In ÉIST ("to listen"), we will be targeting all of these cohorts, but there are choices to be made as to the priorities in developing the initial resources. The most basic requirement is in our view to provide a facility that works for the native speaker communities – regardless of which dialect. For that reason, the system described here, and the resources on which our research has been focussed to date, is geared primarily to native speaker speech.

In testing the present system, test materials were used that contained both native-speaker (L1) speech and L2 speech (the latter from the Mozilla Common Voice[3] collection for Irish). This provides some indicators as to the likely performance of the current system with L1 and L2 speakers and provides some pointers for future work.

3. Resources

The initial efforts to build the ÉIST system drew heavily on the linguistic resources developed for synthesis.

3.1 Speech Corpora

The speech corpora recorded for the synthetic voices were a starting point for the system. These were quite extensive (c.25.2 hours) but involved only 8 speakers. The recordings were of the 3 main dialects referred to above and were based on readings of materials appropriate for each dialect. The quality of recordings was high, and the corpora were edited, cleaned, annotated and aligned – ready to be used in the ASR engine.

Additionally, speech corpora were collected, as part of an initiative MíleGlór[4] ("A Thousand Voices"). A platform was developed that can be used for live or crowdsourced recordings: given that our priority was to obtain data for native speakers of the different Gaeltachtaí (Irish speaking areas), the platform offered different materials, depending on the dialect of the speaker. The control over the text presented to users for recording is important. Ideally, we would like coverage of the sounds in all environments, and it is important to use natural, dialect-appropriate and relatively simple language to ensure it is easy for users to read. Most of the data collected was recorded live during successive Oireachtas gatherings (annual Irish language festival). This corpus is 20.8 hours in duration from 256 speakers reading from dialect-appropriate texts. Prior to recording, demographic information on the speakers is elicited e.g. whether they are native speakers, their dialect, approximate age etc.

We also had access to a corpus of spontaneous speech from 71 speakers, the Comhrá corpus (Uí Dhonnchadha et al., 2012). About 5.1 hours from this corpus has been edited and processed for recognition training, although only part of this is used in the system described below.

Finally, part of the Mozilla Common Voice corpus of Irish (54 speakers, 2.3 hours) was used as part of the Test set for system evaluation (see Section 6). Note that this corpus is of nearly all L2 speakers, and this is something we return to below.

3.2 Lexicon Building

ABAIR synthesis resources were used for building the pronunciation lexicons for the ÉIST system. These are the letter-to-sound (LTS) rules and the pronunciation lexicons for the dialects. The pronunciation lexicons for the text-to-speech systems are rather limited, as they are intended solely to cater for those irregular word forms, whose pronunciation cannot be predicted using the LTS rules. Nonetheless, combined, these provided ideal tools for constructing pronunciation lexicons.

As a strategy for dealing with the multiple dialects in building the synthesis systems, we developed a *Trans-dialect (Trans)* set of LTS rules, which capture the common core of the phonological system, while allowing for dialect-specific modules to capture dialect-specific differences in realisation (Ó Raghallaigh, 2010). From the *Trans* ruleset, we developed a *Trans* lexicon. We also had entirely separate sets of LTS rules for the three dialects, allowing us to build a *Multi-dialect (Multi)* lexicon comprising all forms in all dialects. These two approaches are tested in Section 6.

3.3 Text resources for language modelling

The text corpora used for language modelling included the *Corpus of Irish for Lexicography* (Ó Meachair, M. J. et al., 2021) using the 2021.1 version. It was developed by Gaois, DCU, with funding from Foras na Gaeilge, is referred to as Text A (72m words, 1.5m vocabulary). A version of the *National Corpus of Irish*, provided by Foras na Gaeilge, is referred to as Text B (52m words, c.0.25m vocabulary). The text from a spontaneous speech corpus of Irish is used and is referred to as Text C (c.4m words, c.0.08m vocabulary). Finally, Irish language text collected from Wikipedia is referred to as Text D (c.2.5m words, 0.13m vocabulary)

4. The current ÉIST ASR system

The ASR system is a hybrid system, in that it combines (1) an acoustic model, (2) a pronunciation lexicon and (3) a language model in a weighted finite state transducer (Mohri et al., 2002). We are continuously running experiments making use of various combinations of our speech data for acoustic model training as well as different configurations of lexicons and of the language models. The system described and tested here is built as follows:

4.1 Acoustic Model

The HMM-based neural network acoustic model is a Time-Delay Neural Network (TDNN) (Peddinti et al., 2015; Povey et al., 2018), that was trained on a subset of our speech corpora. This subset was balanced for the 3 dialects and totalled 37.2h, from 281 speakers. 85% of the total speech duration involved native (L1) speakers. Details of the training data are in Table 1. All experiments are done using the Kaldi toolkit (Povey et al., 2011).

[3] https://commonvoice.mozilla.org/

[4] https://abair.ie/mileglor/

Table 1: Details of speech datasets used. Duration is noted in hours.

dataset	#wav	#spk	#words	#vocab	#dur
Train	39,609	281	338,643	15,018	37.24
Test	1174	20	8224	2103	1.14

The data was initially aligned using a triphone GMM-HMM trained using MFCC features, applying linear discriminative analysis (LDA), maximum likelihood linear transformation (MLLT), feature space maximum likelihood linear regression (fMLLR) and speaker adaptive training (SAT). The features for training the TDNN model were 40-dimensional high-resolution MFCCs stacked with 100-dimensional online extracted i-vectors.

Two common, on-the-fly data augmentation techniques were used in training to augment the speech data: *speed perturbation* (Mubashir et al., 2013) and *spectral augmentation* (*SpecAug*) (Park et al., 2019). On-the-fly methods work by augmenting data during training, which both improves the flexibility of training and greatly saves disk space. Using speed perturbation, the training data was tripled using speed warping factors of 0.9, 1.0 and 1.1. *SpecAug* augments the log mel-spectrogram of an utterance, by randomly masking bands on the frequency domain and time domain. This method leads to impressive improvements.

The TDNN model consists of 13 factorized TDNN (TDNN-F) layers with a size of 1024 and a bottleneck size of 128 and was trained for 10 epochs. It was trained with lattice-free maximum mutual information (LF-MMI) (Povey et al., 2016).

System fusion is a common method to make use of multiple similar systems and achieve a stable performance. As the number of training epochs affects how much a system is fine tuned to the training data and as such, how robust it will be to unseen testing data, fusion of variants of acoustic models, trained using a different number of epochs is explored in the evaluation (Section 6), where it is compared to systems trained with a single acoustic model.

4.2 Lexicon

The complexity of pronunciation variation across dialects and speaker communities in Irish is a challenge when developing a pronunciation lexicon. As mentioned above, two different approaches to lexicon building were tested in the present system, a *Trans* lexicon, based on the *Trans* LTS rules, and a *Multi* lexicon, which simply included all dialect pronunciations. This *Trans* lexicon is more compact than the *Multi* lexicon, which has advantages in the size of the decoding lattices and the efficiency with which they can be searched. It would thus confer many advantages if it can perform equally well as the larger *Multi* lexicon. See Table 2 for details.

Table 2: Number of phones / abstract units (#phn) and entries (#lex) in Multi and Trans lexicons.

	Trans	Multi
#phn	92	118
#lex	540k	1006k

4.3 Language Model

The language model in the present hybrid system is a 3-gram model (Goodman, 2001) trained on all text corpora listed in Section 3.3 using the SRILM toolkit (Stolcke et al., 2011). Lattice-rescoring (Liu et al., 2016; Xu et al., 2018) using recurrent neural network language models (RNNLM) (Bengio et al., 2003; Mikolov et al., 2011) has been shown to be greatly beneficial. An RNNLM trained on Text A and Text D is used to rescore the hypotheses generated from the 3-gram language model.

5. System Evaluation

5.1 Test set

A test set was developed for the system evaluation. This consisted of materials taken from two sources. Firstly, a subset of speakers was taken from our own MíleGlór recordings, ensuring no overlap of speakers or utterance text between the training set and the test set. These speakers were virtually all native (L1) speakers. Secondly, part of the Mozilla Common Voice corpus for Irish was used. The data chosen were those where the speaker had declared a dialect preference and predominantly positive listeners' judgements were obtained. These speakers were however L2 speakers, which may partly be explained by the fact that a large cohort of the Irish-speaking online community are not native speakers. Over the two combined sets, efforts were made to balance for the dialects.

The fact that the two datasets used in the test data represented a clean L1/L2 divide is interesting in that it allows inferences to be drawn regarding the likely performance of the ÉIST ASR system for native and non-native speakers. See Table 1 for details of the test set.

5.2 Results

Table 3 presents the results obtained for the *Multi* and *Trans* lexicons. In a), the Overall Word Error Rate (WER) results are compared for: single systems, which are trained for 10 epochs using the baseline 3-gram LM; fused systems using the baseline 3-gram LM; and fused systems rescored using an RNNLM (see Section 4.1). The best results were obtained with the RNNLM, and the *Trans* lexicon performs as well as, or marginally better than the *Multi* lexicon.

Table 3: WER% for Multi and Trans lexicons. a) Overall WER% for all test speakers; b) breakdown of WER according to dialect affiliation of speakers; and c) breakdown of WER% for the two corpora used in the Test set.

a)	Multi	Trans
Single	13.1	13.38
Fused	12.6	12.5
+RNNLM	8.85	8.78
b)	Multi	Trans
Co spk	10.35	9.98
Mu spk	6.96	7.31
Ul spk	10.11	9.74
c)	Multi	Trans
MíleGlór	6.38	6.73
Mozilla	10.68	10.30

In part b) of Table 3, a breakdown of the Overall WER of systems with RNNLM rescoring is presented according to the dialect affiliation of the speakers. It should be noted that the dialect affiliation here refers to the actual dialect in the case of the native speakers (L1) but refers simply to the dialect preference of the L2 speakers, whose speech may approximate to dialect norms in varying degrees. The *Trans* lexicon yields a better performance for Co and Ul speakers, but the *Multi* performs better for Mu.

In part c) of Table 3, WER are compared for the two different corpora used in the Test set, the MíleGlór and Mozilla data. This comparison is of interest because in the former, native speakers dominate (80%) while in the latter all are L2 speakers. There is a consistently large WER difference, with performance being considerably better (lower WERs) for the MíleGlór speakers. This does suggest that the ÉIST system performs better for native-speaker speech. This is not surprising, as this was the intention behind the strong focus of native-speaker speech in the collection of data for the ASR training.

6. Current and future directions

The ÉIST system is available to try[5], although it is still a work in progress. Our current efforts and future aspirations include the following: Speech Corpus extension- we will be gathering a much larger corpus of speech data, focusing on a) Gaeltacht-based native speakers to include all the dialects and b) non-Gaeltacht speakers of Irish. The corpus will be collected in such a way that the different cohorts can be identified, both in terms of native / non-native distinction and the dialect of the speaker. We are currently extending the dialect-appropriate text materials used in MíleGlór for recording, to include much more varied sentences with greatly increased vocabulary.

Further to the current approach, we are also investigating the potential of End-to-End ASR systems (Gulati et al., 2020; Zhang et al., 2020) for dealing with the large variation in Irish speech, including the use of pretrained models, such as Wav2Vec 2.0 (Baevski et al., 2020).

Although not a current activity, we are keenly aware of the need to cater for children's speech, both for synthesis and recognition. This is particularly critical given that much of ABAIR's focus is on developing applications to support Irish language education and to ensure that those with disabilities are included in the Irish language educational, social and cultural spheres.

7. Acknowledgments

This work is part of the ABAIR initiative, which is supported by *An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán*, with funding from the National Lottery, as part of the *Stráitéis 20 Bliain don Ghaeilge*.

8. Bibliographical References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems, 2020-Decem*, 1–12.

Barnes, E., Morrin, O., Ní Chasaide, A., Cummins, J., Berthelsen, H., Murphy, A., Nic Corcráin, M., O'Neill, C., Gobl, C., & Ní Chiaráin, N. (2022). AAC don Ghaeilge: the Prototype Development of Speech-Generating Assistive Technology for Irish. *Proceedings of the International Conference on Language Resources and Evaluation*.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research, 3*(6), 1137–1155.

Ní Chiaráin, N., Comtois, M., Nolan, O., Robinson-Gunning, N., Sloan, J., Berthelsen, H., & Ní Chasaide, A. (2022). Celtic CALL: Strengthening the Vital Role of Education for Language Transmission. *Proceedings of the International Conference on Language Resources and Evaluation*.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*(4), 403–434.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Interspeech 2020*, 5036–5040.

Liu, X., Chen, X., Wang, Y., Gales, M. J. F., & Woodland, P. C. (2016). Two Efficient Lattice Rescoring Methods Using Recurrent Neural Network Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(8), 1438–1449.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., & Černocký, J. (2011). RNNLM --- Recurrent Neural Network Language Modeling Toolkit. *Proceedings of ASRU 2011*, 1–4.

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69–88.

Mubashir, M., Shao, L., & Seed, L. (2013). Audio Augmentation for Speech Recognition Tom. *Neurocomputing*, *100*, 144–152.

Ó Meachair, M. J., Ó Raghallaigh, B., Bhreathnach, Ú., Ó Cleircín, G. & Scannell, K.. (2021). Corpus Creation for Lexicographical Research: Corpas Foclóireachta na Gaeilge (CFG 2020). *Teanga: The Journal of the Irish Association of Applied Linguistics.*, *28*, 278–305.

Ó Raghallaigh, B. T. C. D. (2010). Multi-dialect phonetisation for Irish text-to-speech synthesis : a modular approach. *Sciences-New York, September*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A

[5] https://phoneticsrv3.lcs.tcd.ie/rec/irish_asr

Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, *2019-Septe*, 2613–2617.

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A Time-Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 2–6.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. *Interspeech 2018*, *2018-Septe*(2), 3743–3747.

Povey, D., Ghahremani, P., & Manohar, V. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI Transfer Learning for ASR View project Speech Recognition View project Purely sequence-trained neural networks for ASR based on lattice-free MMI. *Interspeech*, 2751–2755.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *IEEE Signal Processing Society*.

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at Sixteen : Update and Outlook. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 5–9.

Uí Dhonnchadha, E., Frenda, A., & Vaughan, B. (2012). Issues in Designing a Corpus of Spoken Irish. *LREC-2012: SALTMIL-AfLaT Workshop on "Language Technology for Normalisation of Less-Resourced Languages"*, 1.

Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., Carmiel, Y., Povey, D., & Khudanpur, S. (2018). A Pruned Rnnlm Lattice-Rescoring Algorithm for Automatic Speech Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2018-April*, 5929–5933.

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020). Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2020-May*(3), 7829–7833.